Augmenting building energy usage models with data segmentation

A. William Mounter [1], B. Nashwan Dawood [2], C. Huda Dawood[3]
[123]Teesside University, Stephenson St, Tees Valley, Middlesbrough TS1 3BA, United Kingdom

ABSTRACT: *Energy is the lifeblood of modern civilisation, with buildings and building construction contributing to roughly 40% of the global energy usage and CO2 pollution. Predicting building energy consumption is essential for energy management and conservation; data driven models offer a practical approach to predicting building energy usage. The aim of this paper is to improve the data driven models available to aid facility managers in planning building energy consumption.*

In this case study the 'Clarendon building' of Teesside University was selected for use in using it's BMS data (Building Management System) to predict the building's energy usage. With a particular focus on how data segmentation impacts a model's accuracy and computational time, in predicting temperature related building energy use. Specifically, the effect of segmenting data to accommodate seasonality, as well as building activity and dormancy periods. With each data segment to be used to train an ANN model (Artificial Neural Network),to address the different patterns and trends present in each period/segment, using ensemble models where data segmentation overlapped.

The potential of these models were compared on the grounds of accuracy to each other, then discussed to identify the various impacts of segmenting the data. This study was performed as part of a larger study, in improving building energy use predictions during the operational period, by incorporating predicted user behaviors.

KEYWORDS: *Buildings, Neural networks, Data segmentation, Energy, Prediction.*

1. Introduction

The aim of this conference paper is to investigate if data segmentation can be used to improve the prediction of building HVAC energy usage. Data segmentation being the process of dividing and grouping data based on chosen parameters, in this case timeframes, so that it can be used more effectively; (as opposed to data splitting, in which data is randomly split for cross validation usage).

To use an analogy, in cars, winter and summer tyres tend to perform better in their respective seasons than each other and all-season tyres, but poorer than each other and all-season tyres outside of their respective seasons. Would, in the case of machine learning, a model trained with only a season's recorded building data be more accurate at predicting said season's building energy use than a model trained with the variety of data from multiple seasons?

To investigate aim the Clarendon Building, part of Teesside University Campus, was selected for use in this study- due to the data rich environment it's BMS (Building Management system) provided. Previous studies into this building utilising square regression analysis typically had a baseline of "5% Mean Absolute Prediction Error (MAPE)" for the demands of each assets in one day ahead forecasts (Boisson et al.2019). However, the predictions lost accuracy as the rolling horizon increased.

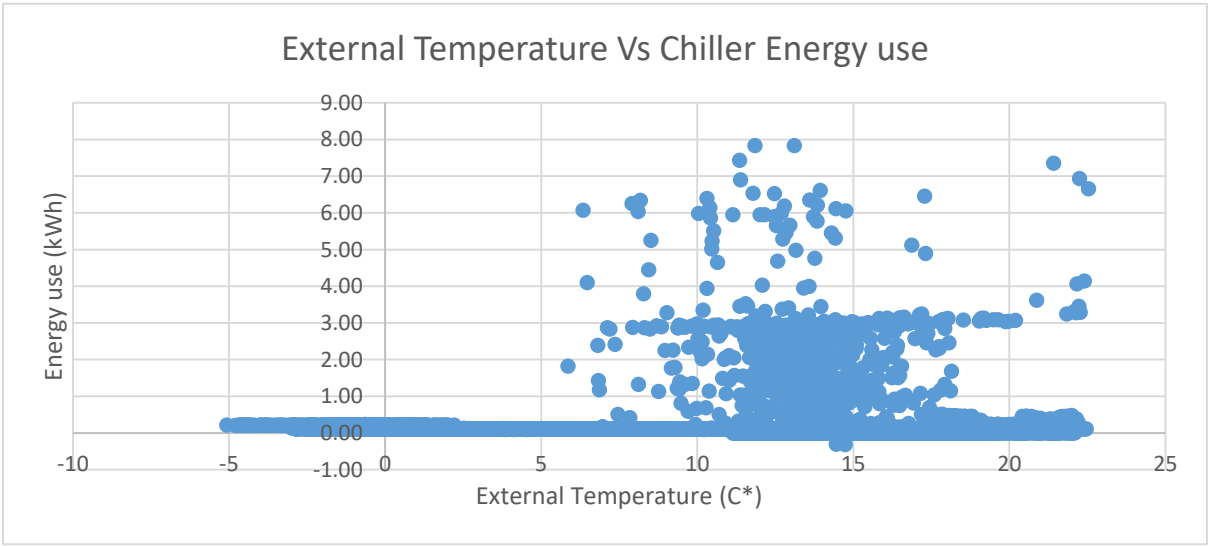
Figure 1: The Clarendon Building, Teesside University (Preston, 2019)

2. Modelling building energy usage

Predicting building energy usage is often complicated by the human element of the process, with no practical and scalable method to capture the personalised energy use information of each energy user in the workplace; 'since the conventional methods use plug-in power meters that are extremely expensive and difficult to maintain over long period of time' (Rafsanjani H, Ghahramani A. 2017). In this case the scope of 'building energy usage' was reduced to part of the Clarendon's HVAC system, specifically the chiller system.

One would expect there to be a positively correlated relationship between the energy used by the building's chiller systems and the external temperature, as the weather gets hotter, the chiller uses more energy to achieve the desired internal temperature. However as Figure 2 demonstrates, this is not the case in an active university building:

Figure 2



This is due to the multiple other factors that impact the building's HVAC usage, than just the external environmental conditions it is trying to shift away from. The internal temperature can be impacted by many sources other than the external temperature, from the number of people in the building, number of active computers to number of open windows, each effected how much work the chiller system has to do. The difference between the internal temperature and what HVAC system is expected to meet would seem like a

reasonable trend that could be correlated to predict the HVAC usage. But once the build reaches the appropriate temperature, the HVAC system will still be expending energy.

Relying upon building's previous average energy usage is a simplistic approach to accommodate for these influential factors, but incorporating them produces more accurate predictions (Wang Z, Srinivasan R. 2017). Two of the main limiting factors in developing building energy use models is integrating occupant behavior into consumption models as well as extending those energy consumption predictions into the long term (Amasyali k, Nora M.2018). The further into the future, the increasingly likely the factors that impact the building energy usage will be different from the ones used to train the model.

By using data segmentation, it may be possible to better accommodate the changing patterns and relationships between the factors that influence the buildings energy usage through creating multiple models, rather than creating a single model can accommodate all of the building different behaviors. Historically data segmentation has been used in marketing to better predict the success of marketing to different groups, developing differing models for potential customers based upon their demographics, lifestyle, behaviors and value (as a customer) (Experian, 2012).

Examples of segmenting building data to improve the accuracy of their prediction models have been met with mixed success. Removing outliers in building energy use data reduced the overall accuracy of its predictions compared to predictions based on unsegmented data in 80% of predictions in one case study (Huyen D, Cetin k, 2018). Whilst in another in the context of energy usage of event venues, found that its models were unable to accommodate both on and off events days, instead requiring separate models to cope with the difference of the behavior of energy usage (Grolinger K, 2016). Depending upon how the data is segmented, their lies the potential to both improve and reduce the accuracy of its predictions.

To test the potential of data segmentation to improve predictions, a method of modeling and predicting building energy use would have to be selected. The most applied machine learning techniques in the field of modelling energy use were: Neural Networks (ANN), Support Vector Machines (SVM), Distribution regression and clustering. (Seyedzadeh, S. 2018). Of these ANN (Artificial Neural Networks) were selected for use for modelling within this study. This was due to ANNs ability to interpret non-liner data compared to other machine learning methods such as multiple linear regression (Which interprets non-linear data poorly) (Zeyu, W & Ravi, S. 2015). Or in the case of Support vector regression, which is also capable of interpreting non-linear data in irregular energy usage environments, due to the size of the datasets available. SVR possessing greater accuracy in smaller datasets than ANN, but being out performed by ANNs in larger datasets (Grolinger, K, Et al, 2016).

ANNs are based upon the concept of establishing a relationship between independent and dependent variables, though the use of training algorithms (Abbas et al, 2019). These relationships are formed in 'hidden layers' in-between the inputs and outputs of the neural networks, where in the equations that assign values to the inputs are systematically randomised dependent on the training algorism used. During the training process the model is fed historical inputs and outputs, of which the training algorithm uses to calibrate the hidden layers continually until its accuracy ceases to increase.

A choice when implementing ANN is whether they are 'feedforward' NN or 'feedback' NN. In Feedforward ANNs, there are no feedback (loops); e.g. the output of any layer does not affect that same or previous layer. In Feedback ANN, networks can have signals travel in both directions by introducing loops in the network (A. Rethinavel Subramanian, 2014). Due to not wishing to overcomplicate the model with accommodating for the feedback loop between the chiller's energy use and the internal temperature or the desired temperature, this case study opted to use a feedforward ANN. This feedforward ANN could then be trained using a levenberg-marquardt training algorithm, due to this method previously having produced the least percent error comparatively in previous investigations of this specific data (Mounter, W. 2019).

3. Research Method

From the Clarendon building, datasets were available of BMS data from October 2017 to August 2019. These datasets contained 15-minute averages of building elements energy usage, as well as sensory data of the internal and external environmental temperatures.

Of these building elements, the building chiller system was selected for use in modelling due to the impact seasonality would have on the overall usage of the chillers. The chillier energy usage was mapped relative to

time, internal and external temperature over the two year period, and segmented. The segments breaking the data into years, seasons, months, weeks, days, weekends, weekdays, building active periods and building dormancy periods.

Following this, using the unsegmented data as a control, the optimum number of hidden layers for the neural network were to be tested for. Too few and the ANN would be too linear to predict the outputs, too many and the ANN would overfit the model. As this was not the central point of the investigating, the number of hidden layers would be tested logarithmically with each's impact on the mean square error and computational time being compared. The number of hidden layers with the least mean square error with a 'reasonable computational period' as determined by the operator will then be used for testing purposes.

The segmented data could then be used to predict the building's chiller's energy usage. In that when is it x temperature externally, how much energy would the chiller's use to achieve y internal temperature, with the mean percent error and computational time of each being recorded. Specifically the following points were be tested:

- The impact of size of training data to a set prediction period, e.g. how a year data set compares to a month data set in predicting the following day's energy usage.
- The impact of size of training data to size of predicted data, e.g. how a year data set compares to a month data set in predicting a range of following periods' energy usage.
- The impact of size of the training data in combination with the period between the training data on the predicted event, e.g. to predict a month period, is it better to use the previous month data to train a model, or the same month in the previous year.
- The impact of segmenting into building active and inactive periods, where in separate models are created for the 8:00am to 18:00pm active period and the 18:15pm to 7:15am period of the data set.
- The impact of segmenting into week days and weekends, where in separate models are created for the Monday to Tuesday period and the Saturday to Sunday period of the data set.

At which point the results of each test could then be compared and reviewed to determine impact of data segmentation on accuracy and computation time

3.1 Method limitations

As the external temperature training data used the temperature at the time of each event, opposed to what the external temperature was predicted to be before the event, this would represent an absolute ideal situation. Where in predicting true future events, the difference between the accuracy of the predicted temperature would impact the overall prediction of the building's energy usage and predicting one year into the future in this manner would be significantly more inaccurate. Though some percentage error could be potentially reduced through greater focus being given to optimise the number of hidden layers of the NN.

4. Results and Discussion

4.1 Computational time

Whilst evaluating the optimum number of hidden layers it was observed that the computation time did not exceed 1 second until the number of hidden layers exceeded 1000 regardless of size of the dataset used. As changing the size of the dataset did not visibly affect the computational time of the process up until 1000 hidden layers, it can be assumed that the number of times the data is processed has a more significant impact on the computational time than the size of the dataset itself. For all following data 10 hidden layers were used, due to the comparably less percent error observed.

Figure 3

Number of hidden layers	10^0	10^1	10^2	10^3	10^4
Percent Error (%)	0.45	0.41	0.42	0.49	28.70
Computation time (Hour, Min, sec)	0.00.01	0.00.01	0.00.01	0.03.20	0.02.34

4.2 Segmenting by period

The following is a selection of the percentage errors observed:

Figure 4.1

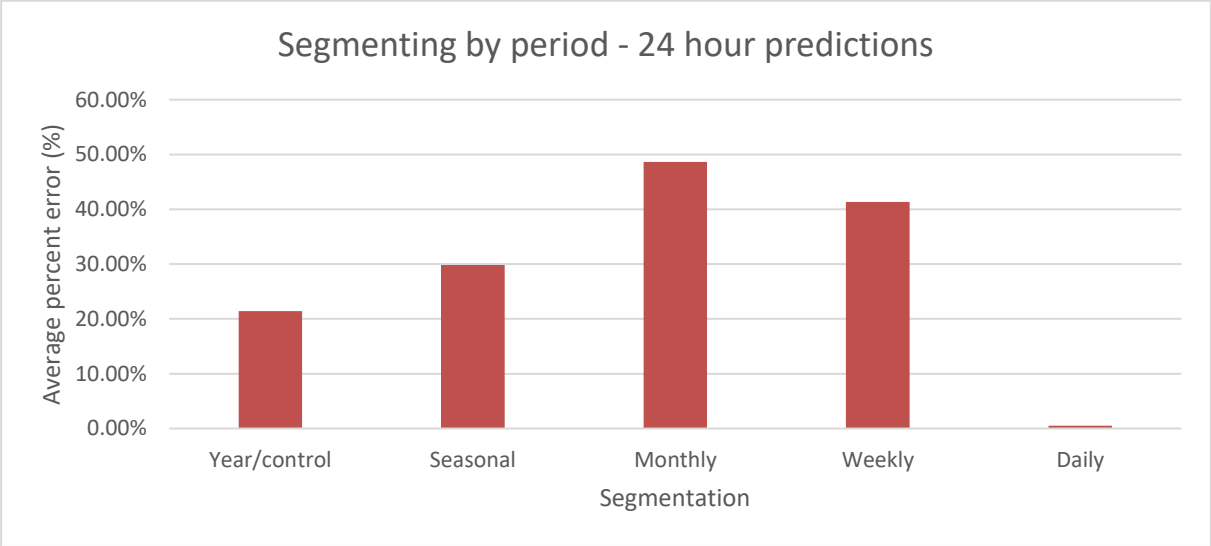


Figure 5.2

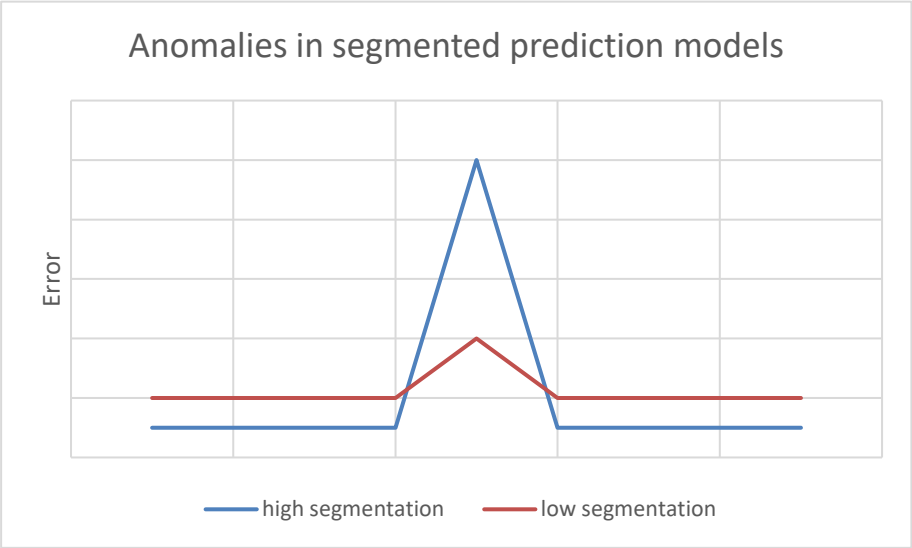
	Yearly	Seasonal	Monthly	Weekly	Daily
Average percent error	21.41%	29.86%	48.62%	41.32%	0.50%

The above is a graphical representation of the average percent error of each segment size after being used to predict the following twenty four hours after the training period. The strange observation was made that the smaller and larger the training datasets, the more accurate on average it's model's predictions would be. Analysis of this trend identified two underling patterns:

- When the training data was highly segmented, that the majority of predictions would increase in accuracy, but the error in predicting 'anomalies' or events which parameters exceeded the training data significantly increased.
- When the training data was lowly segmented, that the majority of predictions would decrease in accuracy relative to highly segmented predictions, but the error in predicting 'anomalies' or events which parameters exceeded the training data would occur less and be less significant.

This these patterns are illustrated and simplified in Figure 5:

Figure 6



The assumed cause of this effect, is that whilst segmenting data makes it's easier to isolate and predict the trends of that period, it increases the difficulty in identifying and modelling underlining trends present in the unsegmented data set as a whole. As the smaller datasets are less likely to have the variety of events occur within its parameters, they will be able to better optimise towards the one's it contains, but unable to account for the ones they do not.

It was additionally observed that using segments data of periods smaller than the predicted period, would significantly reduce the accuracy of the prediction, whilst using segment sizes larger than the predicted period produces the previous effect. Furthermore, the greater the period of time between the training data and the predicted period, the greater error caused by segmenting the data. Noting an exception in that once time between the training data and the predicted period approached that periods position in the previous year, the error would decrease. With predicting seasonally and yearly periods having a higher accuracy on average using their own previous year's data, than the period directly previous to them.

4.3 Segmenting by 'Building Activity and Dormancy'

Due to the monthly segments producing the most error off all prediction periods, they were selected to investigate the potential to isolate internal patterns within segmented periods, as they would demonstrate the largest potential for reduction of in error.

Figure 6.1

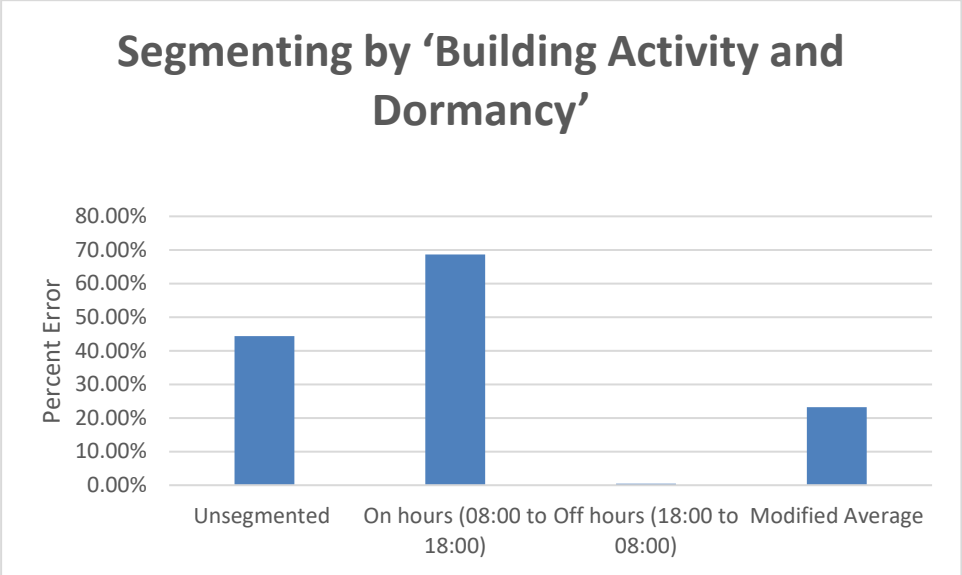


Figure 6.2

	Unsegmented	On hours (08:00 to 18:00)	Off hours (18:00 to 08:00)	Modified Average
Percent Error	44.33%	68.70%	0.48%	23.22%

As shown in figure 6, segmenting the monthly predictions by building activity and dormancy periods, reduced the mean percentage error by 21.10%, from 44.33% to 23.22% error. (Within this context 'modified average' refers to taking the average of both periods with consideration to the different durations of both). The building dormancy period follows a pattern of using 0.1 or 0.2 Kwh depended upon if the external temperature is above or below 0*c respectively, allowing significantly more accurate modelling due to its simplicity than when incorporated with the whole day cycle. Conversely the average mean percentage error of the active period's prediction are increased by 24.37% to 68.70%, indicating that either:

- Removing the night cycle increases the error of predicting the day cycle.
- Removing the nigh cycle does not increase the error of the predicting the day cycle, and that period errors were being 'balanced' by the dormancy period's low errors when the mean error was originally calculated.

Investigation of the specific predictions (rather than means) indicated that it is that later rather than former, with segmenting active and dormancy periods reducing the average mean error of the monthly predictions of the chiller system (under ideal circumstances) by roughly half.

4.4 Segmenting by 'Week days and Week ends'

Given the success in reducing predicting error by separating trends in the data set, it was expected that firstly the weekend and weekdays of a university building would have significantly different usage patterns and thus chiller requirements. And secondly that segmenting along these lines could produce reductions in error similar to segmenting building dormancy and activity periods.

Figure 7.2

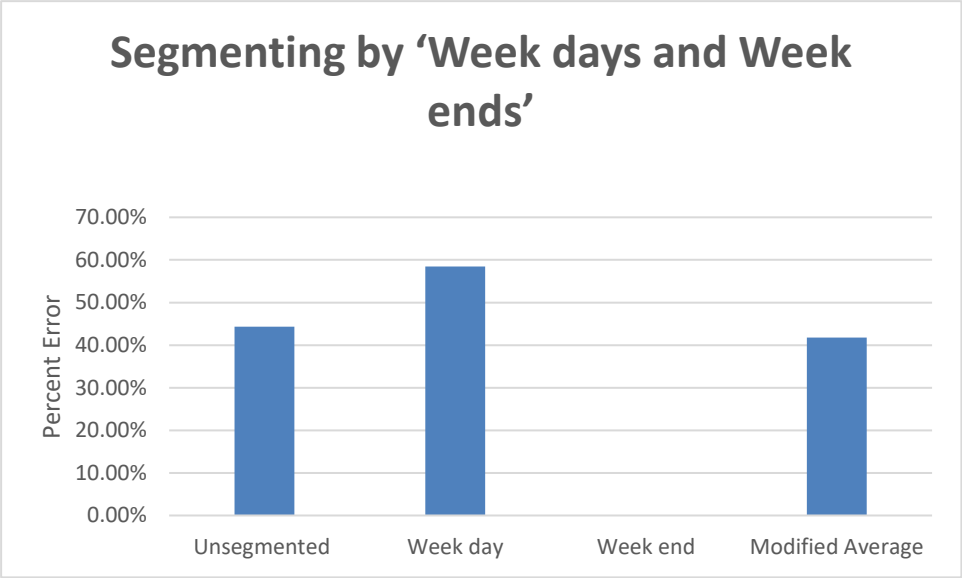


Figure 7.2

	Unsegmented	Week day	Week end	Modified Average
Percent Error	44.33%	58.43%	0.02%	41.75%

As shown in Figure 7, segmenting by weekend and weekday only produced an average reduction of mean percent error by 2.58%. With the weekend period sharing the behaviors of the dormancy period and the active period (but predominantly the former) during the day, and the dormancy period during the night; when active periods did occur during the day, there was an expectation that energy use to be underestimated during the active periods. Though the reduction of error the weekend period was greater than expected; being only 2/7s of the overall prediction period, was not enough to significantly counter balance the increase in error observed in the means of the remaining 5/7s of the period (relative to active and dormancy periods).

4.5 Segmenting by both 'Weekdays and weekends' and 'Building Activity and Dormancy'

Combining the two processes together, and segmenting monthly data into four parts, produced the average percent errors shown in figure 8:

Figure 8.1

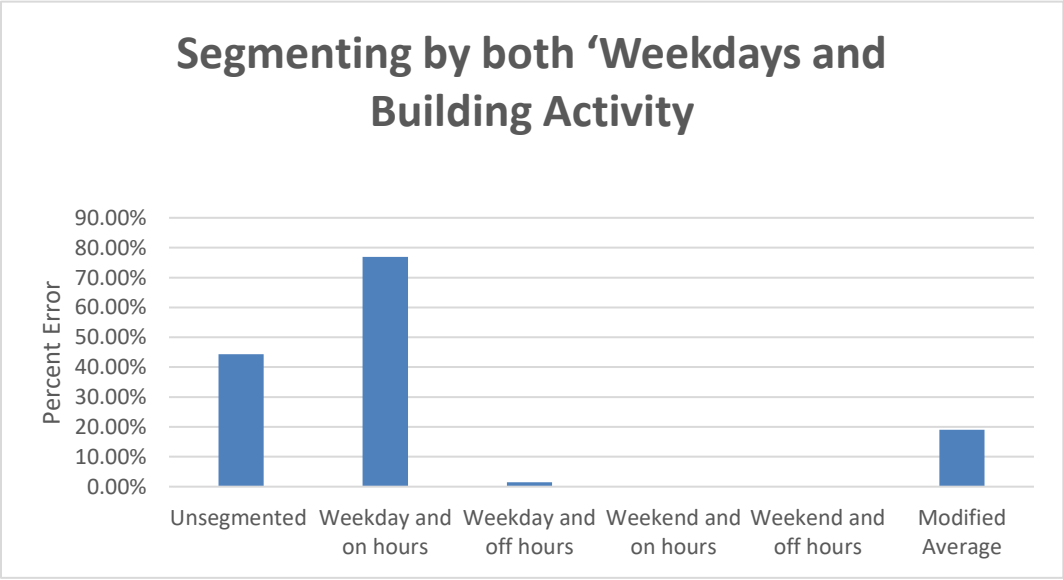


Figure 8.2

	Unsegmented	Weekday and on hours	Weekday and off hours
Percent Error	44.33%	76.93%	1.48%

Figure 8.3

	Weekend and on hours	Weekend and off hours	Modified Average
Percent Error	0.01%	0.05%	19.03%

Doing so reduced the average of monthly prediction's percent error from 44.33% to 19.03%, reducing error by 25.29%, a further 4.19% reduction in error relative to segmenting by dormancy and active periods alone. At this stage combining the weekend dormancy period with the weekday dormancy period to simplify the system into three rather than four models, Would likely not significantly impact the overall accuracy of the system, but as there is a higher chance for building activity after 18:00pm during the weekdays compared to the weekend, it is possible that the differences between the two periods could minimally increase the average percent error relative to them being modelled separately.

On average, in situations where the predicted period is expected to be similar to the training data, segmenting appropriately to the patterns present in the data, without reducing the net data used to below the duration of the predicted period produced the least error. Whilst in situations where the predicted period is expected to be dissimilar, or repeatedly/significantly exceed the parameters of the training data, segmenting less with the aim of expanding the training data's parameters to accommodate the outliers produced the least predictions error.

5. Conclusion

In conclusion, data segmentation can have both a negative and positive impact upon the accuracy of predicted building energy usage dependent upon: the duration of the predicted period, the time between the training data and the predicted events as well as patterns of the segmented data. In that:

- The greater the segmentation, the greater the accuracy in predicting none anomalous data; but the greater the error caused by anomies in each segment.
- The more correlated the relationship between the segments and the inputs, the more likely segmentation will decrease prediction error; but less correlated the relationship between the segments and the inputs, the more likely segmentation will increase average error.
- The greater the size of the predicted period and time between the predicted event and training data to the size of the training data the more negative the impact of data segmentation will have on prediction accuracy. Due to the larger the period and further away the prediction, the increasing likely there will be anomies outside of the range of the training data.

Under the ideal conditions of predicting one day into the future, using a one-day segment to train the ANN, with completely accurate temperature data, an average mean percent error of 4% could be achieved. Though it can be expected that this error would increase, in the case of predicting future energy usage based upon predicted weather data for the external temperatures and the building temperate comfort zone for the internal.

Based upon these results, five main areas of future work were identified:

- Investigating the accuracy of smaller data segments such as hours in predicting shorter periods into the future.
- Using predicted weather data to investigate its impact on prediction accuracy, and test the model's robustness, as the actual weather conditions would not be available when predicting the future.
- Investigating the accuracy of other machine learning techniques, such as SVR for use in smaller data segments, or other types of ANN and training algorithms as well as hybridisations of these models.
- Investing the impact of data segmentation using other HVAC systems, as well as predicting the energy usage of the HVAC system as a whole.
- Directly comparing the percent error of this method of predicting energy usage with other methods

previously used on the Clarendon building, using the same data to train both models, such as the square regression model used in 'Boisson et al.2019'.

6. References

ANN models for prediction of residual strength of HSC after exposure to elevated temperature, Fire Safety Journal, Volume 106, 2019, Pages 13-28

A. Rethinavel Subramanian, Int. Journal of Engineering Research and Applications, Vol. 4, Issue 1(Version 2), January 2014, pp.237-241

Hamed Nabizadeh Rafsanjani, Ali Ghahramani, Towards utilizing internet of things (IoT) devices for understanding individual occupants' energy usage of personal and shared appliances in office buildings,Journal of Building Engineering,Volume 27.

Husain Abbas, Yousef A. Al-Salloum, Hussein M. Elsanadedy, Tarek H. Almusallam,

Huyen Do, Kristen S. Cetin, Evaluation of the causes and impact of outliers on residential building energy use prediction using inverse modelling, Building and Environment, Volume 138, 2018, Pages 194-206

Kadir Amasyali, Nora M. El-Gohary, A review of data-driven building energy consumption prediction studies, Renewable and Sustainable Energy Reviews, Volume 81, Part 1, 2018, Pages 1192-1205

Katarina Grolinger, Alexandra L'Heureux, Miriam A.M. Capretz, Luke Seewald, Energy Forecasting for Event Venues: Big Data and Prediction Accuracy, Energy and Buildings, Volume 112, 2016, Pages 222-233

Mauro Ribeiro, Katarina Grolinger, Hany F. El Yamany, Wilson A. Higashino, Miriam A.M. Capretz, Transfer learning with seasonal and trend adjustment for cross-building energy forecasting, Energy and Buildings, Volume 165, 2018, Pages 352-363

Mounter, W. Dawood, H. Dawood, N. (2019). The impact of data segmentation on modelling building energy usage. In The International Conference on Energy and Sustainable Futures (ICESF). Nottingham Trent

preston.(2019). *Tessideuniversity.* Available:http://prestonvsteesside.e-monsite.com/album/teesside-university/. Last accessed 2019.

Pierre BOISSON, Simon THEBAULT, Sergio RODRIGUEZ, Sylvia BREUKERS, Richard CHARLESWORTH, Sarah BULL, Igor PEREVOZCHIKOV, Mario SISINNI, Federico NORIS, Mihai-Tiberiu TARCO, Andrei CECLAN, Tom NEWHOLM,. (2017). *DR Bob D5.1.*Available: https://www.dr-bob.eu/wpcontent/uploads/2018/10/DRBOB_D5.1_CSTB_Update_2018-10-19.pdf. Last accessed 2019.

Seyedzadeh et al. Visualization in Engineering (2018) 6:5

Use prediction: Contrasting the capabilities of single and ensemble prediction models, Renewable and Sustainable Energy Reviews, Volume 75, 2017.

Wang, Zeyu and Ravi S. Srinivasan. "A review of artificial intelligence based building energy use prediction: Contrasting the capabilities of single and ensemble prediction models." (2017).

University, 9th-11th September 2019. Nottingham Trent University: ICESF. 127-133.

Zeyu Wang, Ravi S. Srinivasan, A review of artificial intelligence based building energy

Unknown. (Unknown). Digital Segmentation. *Available: https://www.experian.co.uk/assets/marketing-services/white-papers/wp-digital-segmentation.pdf. Last accessed 18/10/2019*